

· 临床研究 ·

基于数据挖掘与可视化技术的 2 型糖尿病 疾病关联性分析

杨志清¹, 张靓¹, 郭玲玲¹, 郭燕青¹, 郝崇奇¹, 王海星¹, 齐昊²

1. 山西医科大学第一医院医疗大数据中心,山西 太原 030001;

2. 山西医科大学第一医院内分泌科,山西 太原 030001

摘要: 目的 基于数据挖掘与可视化技术构建 2 型糖尿病疾病关联网络,揭示 2 型糖尿病相关疾病关联关系。方法 通过山西医科大学第一医院医疗大数据平台最终整理得到 33 723 条记录,210 861 条有效诊断。利用共词分析方法,构建 2 型糖尿病疾病共现网络,并运用 Gephi 软件进行可视化展示。结果 构建的 2 型糖尿病疾病共现网络包含 442 个节点,2 304 条连线,网络密度为 0.024,平均最短路径为 1.979;经模块化分析,网络聚类为 11 个社区,其中节点数大于 10,词频占比大于 5% 的社区有 6 个,分别为 2 型糖尿病为中心的社区(可细分为高血压/高血脂集群、甲状腺相关疾病集群、骨关节病集群)、高血压及其脑血管病集群、冠心病集群、心律失常集群、2 型糖尿病并发症集群、肺心病及其症状集群,其中 2 型糖尿病与高血压 3 级的关联最密切,网络连线最粗。结论 通过数据挖掘与可视化技术,展示了 2 型糖尿病关联疾病集群,有机会为 2 型糖尿病综合诊疗提供新的思路。

关键词: 2 型糖尿病; 电子病历; 共词分析; 可视化

中图分类号: R 587.1 文献标识码: B 文章编号: 1674-8182(2020)11-1490-05

Association analysis of type 2 diabetes based on data mining and visualization technology

YANG Zhi-qing*, ZHANG Liang, GUO Ling-ling, GUO Yan-qing, HAO Chong-qi, WANG Hai-xing, QI Hao

* Department of Health and Medical Big Data Center, First Hospital of Shanxi Medical University, Taiyuan, Shanxi 030001, China

Corresponding author: QI Hao, E-mail: qi-h@vip.163.com

Abstract: **Objective** To Construct an association network for type 2 diabetes based on data mining and visualization technology, revealing the association relationship for type 2 diabetes. **Methods** Through the Health and Medical Big Data platform of the First Hospital of Shanxi Medical University, 33 723 records and 210 861 valid diagnoses were finally sorted out. The co-word analysis method was used to construct the co-occurrence network of type 2 diabetes, and Gephi software was used for visual display. **Results** The constructed type 2 diabetes co-occurrence network contained 442 nodes, 2 304 connections, a network density of 0.024, and an average shortest path of 1.979. After modular analysis, the network was clustered into 11 communities, of which the number of nodes was greater than 10. There were 6 communities whose word frequency accounts for more than 5%, namely: type 2 diabetes-centered communities (which can be subdivided into hypertension/hyperlipidemia clusters, thyroid-related diseases clusters, osteoarthropathy clusters), hypertension and its cerebrovascular disease cluster, coronary heart disease cluster, arrhythmia cluster, type 2 diabetes complication cluster, pulmonary heart disease and its symptom cluster, among which type 2 diabetes was the closest to hypertension level 3, and the network connection is the thickest. **Conclusion** Through data mining and visualization technology, it shows the clusters of type 2 diabetes related diseases, and has the opportunity to provide new ideas for the comprehensive diagnosis and treatment of type 2 diabetes.

Key words: Type 2 diabetes; Electronic medical record; C-word analysis; Visualization

Fund program: Key Research and Development Project of Shanxi Province (201803D31099)

国际糖尿病联盟(IDF)数据显示,2017年世界成年人口(22~79岁)的糖尿病患病率约为8.8%,而我国作为世界上糖尿病患病人口最多(1.144亿)的国家,患病率高达11.0%,患病人数占世界糖尿病总人数的近四分之一^[1]。根据联合国人口司的数据,考虑人口结构和城市化进展,中国在2045年的成年糖尿病患病人口将达到约1.2亿^[2]。2型糖尿病(T2DM)是机体胰岛素抵抗或胰岛素相对缺乏的一类代谢异常疾病,其患病总数占全部糖尿病患者的90%~95%^[3]。糖尿病及其并发症所造成的疾病负担已成为当前世界各国共同面临的重大公共健康问题和社会经济问题。

随着医院信息化进程的不断加快,医院信息化系统中积累了海量的数据信息,这些数据隐藏了很多有价值的信息,如何通过数据挖掘技术使这些数据辅助医疗机构科学决策,使数据价值最大化,成为当前研究热点。共词分析方法是数据挖掘方法中的一种,其最早在20世纪70年代中后期由法国文献计量学家提出^[4],原理主要是对一组词两两统计它们在同一篇文献中出现的次数,通过聚类分析研究这些词所代表的学科和主题的结构变化。其后,有学者尝试以生物医学实体共现的方法从大规模生物医学文本中探测被隐藏的关联,并取得较好的效果^[5-7],但这些研究多关注生物医学实体词组在文献中的共现。近年来,针对电子病历数据进行挖掘成为研究热点,对电子病历进行分析是一种以患者为中心的研究方法,是了解患者病情的重要手段^[8],已有不少研究以患者电子病历为基础开展自然语言处理、可视化等分析^[9-10],然而基于共词分析可视化技术进行生物医学实体关联分析,挖掘潜在生物医学知识的研究较少。

本研究拟利用共词分析方法,通过有效的数据可视化技术,从电子病历入手,基于共现关系构建2型糖尿病相关疾病共现网络,帮助医学科研人员快速准确地发现隐藏的疾病间的关联,以期为发现2型糖尿病治疗新靶标、针对患者不同疾病开展综合治疗以及新药研发等提供一定的思路,为2型糖尿病的综合诊疗防控提供一定的数据基础。

1 材料与方法

1.1 数据源

数据来源于山西医科大学第一医院医疗大数据平台,该平台将医院既往各个信息系统中的数据进行整合,通过患者唯一编码将患者信息进行关联,并利用自然语言处理、机器学习、ETL工具等技术对数据进行清洗、归一、结构化等处理,形成了可供病

历检索、建立项目、结构化字段提取、数据导出的科研一体化平台。在平台中,以出院诊断包含“2型糖尿病”为检索条件,检索2011年至2017年的住院病例,共得到24208个患者的33789份病历资料,同一患者的不同时间就诊以多条记录保留,提取患者的一般特征及出院诊断等字段。

1.2 数据预处理

利用Excel 2010、SAS 9.4等软件对得到的病例资料进行清理,删除病历资料不完整的记录,并对出院诊断字段按照标准疾病编码进行标准化清理,删除无法归类的无效诊断、归类后同一条记录的相同诊断等,保留上下级诊断以保证结果的精度。最终整理得到33723条记录,210861条有效诊断。

1.3 构建共现矩阵

运用共词分析法进行疾病与疾病的关联分析,共词分析方法能够统计出同一就诊记录中各疾病诊断共同出现的次数,在同一就诊记录中出现的次数越多,则说明这两种疾病的关系越紧密。利用Python自编程序语言,提取诊断字段中疾病与疾病的共现关系,即每一条记录下任意两疾病诊断构成的无序共现对,构建疾病共现矩阵。

1.4 疾病共现网络构建

Gephi是一款基于JVM的开源免费复杂网络分析系统,可用于探索性数据分析、社交网络分析、生物网络分析等,不仅提供基本的中心性分析,而且还提供复杂网络分析功能和动态网络分析功能,网络分析的效率高、功能强、可视化效果好。本研究以电子病历中疾病诊断信息为节点,疾病间的共现关系为边,将上述构建的疾病共现矩阵导入到Gephi软件中,构建2型糖尿病诊断相关共现网络图谱,并对网络进行特征分析。

1.5 疾病共现网络分析

本研究拟从宏观与微观两个层面分析疾病共现网络的结构特征,从宏观层面上对网络的节点数、连线数、密度、最短路径等特征进行分析,从微观层面上,分析网络点度中心度、接近中心度、中介中心度,挖掘疾病与疾病之间的关联关系。(1)宏观指标。**节点:**是网络中的最小单位,一个节点表示一名行动者,在本研究中代表一种疾病。**点度:**是一个节点所拥有的连线数量。本研究中代表疾病与疾病在几条记录中共现。**连线:**网络中两点之间的联系。在社会网络分析中,连线可以代表任何社会关系,连线用它的两个端点来定义。**密度:**指简单网络中实际存在的连线数量,以其占最大可能连线数的比例来表示,它描述了一个网络中各个点之间关联的紧密程度。**最短路径:**在网络中,两个节点之间的“距离”,就是两点之间最短路径所含的连线数,这条最短路径称为捷径(geodesic)。(2)微观指标。**点度**

中心度(Degree, 明星度)：是指与该节点有连线的节点的数量。在一个网络中，如果一个节点与其他节点间存在大量的直接联系，则该节点处于网络的中心地位，在该网络中拥有较大的“权力”。接近中心度(Closeness, 信息传播者)：能反应节点在整体网络中影响力的可扩展性。社会网络分析中，接近中心度用于测量该节点通过网络到达其它节点的难易程度。一个节点与其他所有节点之间的距离越小，信息就越容易到达这个节点，这个节点的接近度也就越高。中介中心度(Betweenness, 信息桥梁)：即所有通过该节点的任意两节点间最短路径的条数，较高的中介中心度说明该行为者作为知识在网络传播的媒介作用明显。

2 结 果

2.1 基本情况描述 本次纳入的 33 723 条就诊记录中，男性 19 292 例次，占比 57.21%，女性 14 430 例次，占比 42.79%。患者年龄 13~100 (63.5 ± 12.7) 岁。2 型糖尿病患者入院时间分布见图 1。2011 年至 2012 年为该院信息系统逐步完善阶段，故在 2011 年保留的完整患者信息较少，另由于医疗大数据平台数据更新至 2017 年 7 月，故曲线在 2017 年呈下降趋势，由分布图可见，在 2012 年至 2016 年间，该院就诊的 2 型糖尿病患者呈增高趋势，符合我国糖尿病患病人数逐年递增的整体趋势。2 型糖尿病患者入院频数前 10 位的科室分布见图 2，可见 2 型糖尿病患者主要就诊科室为内分泌科、心内科、神经内科、老年病科次之，是因本次纳入患者以 60 岁以上老年人居多，患者主要分布在以冠心病、脑梗塞等老年性多发病为主的相关科室。

2.2 2 型糖尿病相关诊断分析 将 210 861 条有效诊断进行频数统计，得到 3 493 类诊断，占比 $\geq 1.5\%$ 的前 10 种诊断疾病见图 3，占全部诊断的 36.9%。与“2 型糖尿病”同时出现次数较高的诊断为高血压 3 级、脂肪肝、高脂血症、高血压 2 级、脑梗死、2 型糖尿病性肾病、冠状动脉粥样硬化性心脏病、不稳定型心绞痛，为 2 型糖尿病的常见伴随诊断。

2.3 疾病诊断共现网络分析 利用 Gephi 绘制疾病诊断共现网络，为了使图片更清晰，选择共现频次大于 40 的诊断共现对构建网络。结果见图 4。图中不同颜色代表不同的社区，圆点越大代表该节点的连线数越多，即与其他疾病共现关系越多，连线的粗细代表共现频次的大小，疾病与疾病共现频次越多，连线越粗。可见整个网络分为明显的几个社区，其中 2 型糖尿病即本次研究目标疾病节点最大，其所在的社区

在网络中分布最广，其余可见高血压、高脂血症、脂肪肝等也均为共现网络中的核心连接枢纽，2 型糖尿病与高血压 3 级的连线最粗。

构建的网络共有 442 个节点，2 304 条连线，网络密度为 0.024，平均最短路径为 1.979；对节点的点度中心度、接近中心度、中介中心度进行计算并排序，选取前 10 位进行对比(表 1)，可见 2 型糖尿病、高血压 3 级、脂肪肝均位于前三，说明这三种疾病在网络中均处于中心位置，对共现网络的影响力均较大，腔隙



图 1 2 型糖尿病患者入院时间分布图

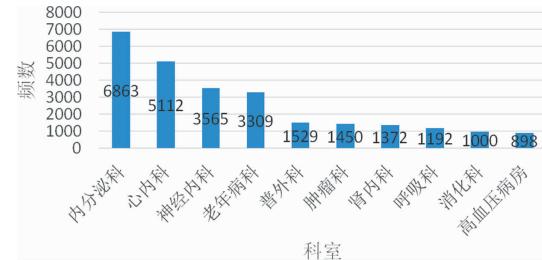


图 2 2 型糖尿病患者入院频数前 10 位的科室分布图



图 3 2 型糖尿病患者前 10 种诊断疾病的百分比分布图

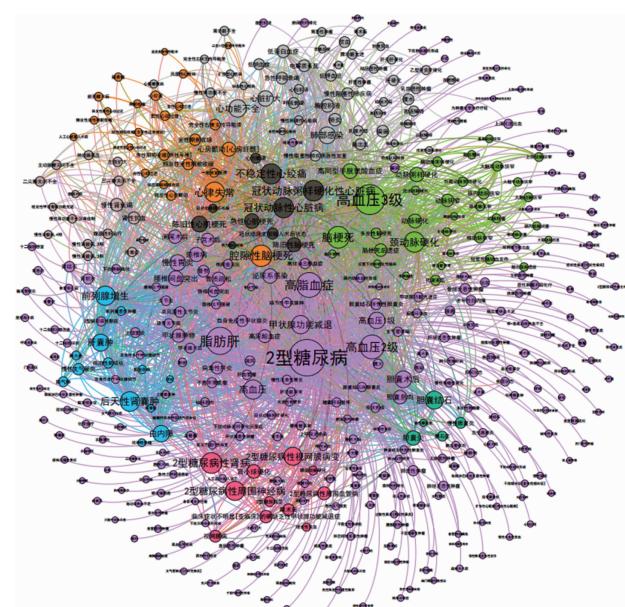


图 4 2 型糖尿病相关疾病共现网络图

性脑梗死的点度中心度及接近中心度均靠前,说明其连接数较多且较易接近,但其中介中心性靠后,表示其对网络的桥梁作用相对较弱;另心律失常、高脂血症、高血压 2 级、颈动脉硬化、脑梗死,均在 2 型糖尿病诊断共现网络处于关键位置的诊断。对疾病共现网络进行模块化分析,共得到 11 个社区分组,社区节点数量大于 10,词频占比大于 5% 有 6 个社区。社区词频占比及节点数见表 2。

表 1 2 型糖尿病疾病共现网络特征指标前 10 位对比

序号	点度中心度	接近中心度	中介中心度
1	2 型糖尿病	2 型糖尿病	2 型糖尿病
2	高血压 3 级	高血压 3 级	高血压 3 级
3	脂肪肝	脂肪肝	脂肪肝
4	腔隙性脑梗死	腔隙性脑梗死	心律失常
5	心律失常	高脂血症	高脂血症
6	高脂血症	心律失常	高血压 2 级
7	高血压 2 级	高血压 2 级	腔隙性脑梗死
8	颈动脉硬化	颈动脉硬化	颈动脉硬化
9	脑梗死	脑梗死	脑梗死
10	后天性肾囊肿	后天性肾囊肿	前列腺增生

表 2 2 型糖尿病疾病共现网络社区前 6 位

社区	颜色	词频	占比	节点数
1	粉	94466	49.00	278
2	绿	26721	13.86	35
3	黑	13877	7.20	10
4	黄	11815	6.13	25
5	红	11761	6.10	11
6	灰	11284	5.85	25

3 讨 论

2 型糖尿病疾病共现网络社区 2(绿色)代表高血压及其脑血管并发症,既往研究显示,2 型糖尿病与高血压有共同的发病基础胰岛素抵抗^[11],故二者在临幊上常相伴出现,约 75% 的 2 型糖尿病患者合并有高血压^[12]。本次分析发现,2 型糖尿病患者与高血压共同出现频次最高,且其与高血压 3 级关系最密切。社区 3(黑色)代表冠心病集群,2 型糖尿病患者体内脂代谢会发生紊乱,且长期高糖状态会通过加剧糖基化与氧化作用使血管张力改变,血压升高,加速动脉的硬化和钙化,另外方慧等^[13]指出,2 型糖尿病患者体内代偿性高胰岛素血症会通过“双信号通路”促进冠心病的发生。社区 4(黄色)代表心律失常集群,2 型糖尿病患者由于体内糖、脂代谢紊乱容易引起微血管病变,引起糖尿病性心肌病,影响心肌传导,再者,2 型糖尿病还易损害自主神经,交感与副交感神经兴奋失衡而引起患者心率及传导异常^[14]。社区 5(红色)代表 2 型糖尿病并发症集群,血管病变是糖尿病患者最常见的并发症^[15],尤其是心血管并发

症,约有 80% 的糖尿病患者死于心血管疾病^[16]。社区 6(灰色)为肺心病及其相关症状,在肺心病失代偿期,由于糖皮质激素的使用,可使胰岛 β 细胞破坏,胰岛素分泌不足,反馈调节使血糖增加,促进糖尿病的发生,肺心病合并 2 型糖尿病是成为加重病情,加速患者死亡的重要原因之一^[17]。

社区 1(粉色)即以 2 型糖尿病为中心的社区,其连接的度大于 10 的节点又可细分为几个小的模块,分别为:高血压/高血脂模块、甲状腺相关疾病、骨关节相关疾病。研究显示,糖尿病患者中甲状腺功能异常的发生率是非糖尿病患者的 2~3 倍^[18],合并甲状腺疾病的发病率较健康人群高 4.8%~31.4%^[19],可能为患者胰岛素分泌受限,导致脱碘酶活性下降,甲状腺激素的浓度下降,机体代谢受损,最终使得碘泵功能障碍,并发各种甲状腺疾病^[20],其他如炎症、基因、免疫等多种因素均可能为其影响因素。众所周知,1 型糖尿病患者易合并自身免疫性甲状腺疾病是与二者具有共同的遗传易感基因有关,而自身免疫性因素在 2 型糖尿病的发病机理中也扮演着越来越重要的角色^[21]。余群^[22]对 221 例 2 型糖尿病患者的甲状腺功能进行回顾性分析显示,T2DM 患者血清 TPOAb 阳性率较非糖尿病患者高,TADM 与自身免疫性甲状腺炎关系密切,本次构建网络中同样发现,2 型糖尿病与自身免疫性甲状腺炎关联较密切,与既往研究一致。

有流行病学证据显示,高血糖可能会导致骨关节炎退行性疾病的进展,美国第三次健康与营养调查数据也显示,糖尿病患者患有骨关节炎的比率几乎为非糖尿病患者的 3 倍^[23]。Nieves-Plaza 等^[24]的研究发现,糖尿病患者的骨关节炎发病率(49%)要远高于非糖尿病患者(26.5%)。Schett 等^[25]在进行 20 年的随访研究后发现,校正年龄、体质指数及其他危险因素后,2 型糖尿病是骨关节炎的独立危险因素,本次网络关系也显示,2 型糖尿病与骨关节相关疾病(骨关节炎、骨质疏松等)关联密切。其可能原因如下。(1)炎症:糖尿病作为代谢综合征的一部分加重了关节炎症反应,促进了骨关节炎的进展;(2)糖基化终末产物 AGEs:体内持续高糖会使糖基化终末产物 AGEs 聚集,其在糖尿病性骨关节炎的发病机制中发挥重要作用;(3)软骨细胞 GLUT 功能异常:糖尿病可干扰软骨细胞葡萄糖转运蛋白 GLUT 的表达和功能,导致胞内葡萄糖蓄积,软骨细胞凋亡,促进骨关节炎发生发展^[26]。另糖尿病患者,在体内长期高糖的作用下,血液渗透压增加,渗透性利尿使得尿中排出的钙磷增多,引起骨质疏松;此外,产生的过多

AGEs 会引起骨胶原蛋白结构改变,影响骨盐沉积,使得骨脆性增加^[27]。

综上所述,本研究利用共词分析的方法,挖掘电子病历中 2 型糖尿病相关诊断信息,通过 Gephi 软件构建疾病共现网络,发现与 2 型糖尿病关联较密切的疾病集群有:高血压及其脑血管病集群、冠心病集群、心律失常集群、2 型糖尿病并发症集群、肺心病及其症状集群、甲状腺相关疾病集群、骨关节病集群等,其中,2 型糖尿病与高血压 3 级的关联最密切,网络连线最粗。本研究利用数据挖掘及可视化技术,展示了 2 型糖尿病相关疾病关联关系,可能为 2 型糖尿病临床诊疗、科研选题等提供一些新的思路,为提高 2 型糖尿病综合防治降低国家糖尿病患病率提供一定的研究基础。但本次研究仅构建了疾病间的关联关系,其关联的生物学机制有待进一步深入研究。

参考文献

- [1] International Diabetes Federation (IDF). IDF Diabetes Atlas [EB/OL]. 8th Edition. Brussels, Belgium: International Diabetes Federation, 2017. [2018-10-15]. <http://www.idf.org/e-library/epidemiology-research/diabetes-atlas.html>.
- [2] Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045 [J]. *Diabetes Res Clin Pract*, 2018, 138: 271–281.
- [3] American Diabetes Association. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes-2018 [J]. *Dia Care*, 2018, 41 (Supple 1) : S13–S27.
- [4] Cozzens SE, Callon M, Law J, et al. Mapping the dynamics of science and technology:sociology of science in the real world [J]. *Contemp Sociol*, 1988, 17 (6) : 815.
- [5] Chun HW, Tsuruoka Y, Kim JD, et al. Extraction of gene-disease relations from medline using domain dictionaries and machine learning [J]. *Pac Symp Biocomput*, 2006; 4–15.
- [6] Zha ML, Cai JY, Chen HL. A bibliometric analysis of global research production pertaining to diabetic foot ulcers in the past ten years [J]. *J Foot Ankle Surg*, 2019, 58 (2) : 253–259.
- [7] Zhang J, Xie J, Hou WL, et al. Mapping the knowledge structure of research on patient adherence:knowledge domain visualization based co-word analysis and social network analysis [J]. *PLoS One*, 2012, 7 (4) : e34497.
- [8] Meng X, Yang JJ. Visual analysis for type 2 diabetes mellitus-based on electronic medical records [M]//Smart Health. Cham: Springer International Publishing, 2014: 160–170.
- [9] 包小源, 黄婉晶, 张凯, 等. 非结构化电子病历中信息抽取的定制化方法 [J]. 北京大学学报(医学版), 2018, 50 (2) : 256–263.
- [10] 石习敏, 陈娟, 杨均雪, 等. 基于知识图谱的国内外医学数据挖掘研究可视化探析 [J]. *中国全科医学*, 2017, 20 (21) : 2623–2628.
- [11] 王连英, 张秀英, 孔祥双, 等. 糖尿病前期与高血压关系及其影响因素分析 [J]. *中华健康管理学杂志*, 2019, 13 (4) : 308–313.
- [12] Colosia A, Khan S, Palencia R. Prevalence of hypertension and obesity in patients with type 2 diabetes mellitus in observational studies: a systematic literature review [J]. *Diabetes Metab Syndr Obes: Targets Ther*, 2013; 327.
- [13] 方慧, 王旭开. 胰岛素在 2 型糖尿病合并冠心病中作用的研究进展 [J]. *中华老年多器官疾病杂志*, 2013, 12 (4) : 318–320.
- [14] 严颖, 陆炜. 2 型糖尿病住院患者发生心律失常的类型及相关危险因素分析 [J]. *心电与循环*, 2017, 36 (2) : 89–92.
- [15] International Diabetes Federation Guideline Development Group. Global guideline for type 2 diabetes [J]. *Diabetes Res Clin Pract*, 2014, 104 (1) : 1–52.
- [16] Lorber D. Importance of cardiovascular disease risk management in patients with type 2 diabetes mellitus [J]. *Diabetes Metab Syndr Obes*, 2014, 7: 169–183.
- [17] 曲认秀, 李凯伟. 肺心病合并 2 型糖尿病临床特点浅析 [J]. *医学信息*, 2015, 28 (2) : 222–223.
- [18] Duntas LH, Orgiazzi J, Brabant G. The interface between thyroid and diabetes mellitus [J]. *Clin Endocrinol*, 2011, 75 (1) : 1–9.
- [19] Kadiyala R, Peter R, Okosieme OE. Thyroid dysfunction in patients with diabetes: clinical implications and screening strategies [J]. *Int J Clin Pract*, 2010, 64 (8) : 1130–1139.
- [20] 高建勤, 张晶, 季宇, 等. 1255 例住院 2 型糖尿病患者甲状腺疾病患病率分析 [J]. *临床内科杂志*, 2011, 28 (9) : 600–602.
- [21] Elebrashy IN, El Meligi A, Rashed L, et al. Thyroid dysfunction among type 2 diabetic female Egyptian subjects [J]. *Ther Clin Risk Manag*, 2016, 12: 1757–1762.
- [22] 俞群. 221 例 2 型糖尿病患者甲状腺功能回顾分析 [J]. *实用糖尿病杂志*, 2016, 12 (4) : 20–21.
- [23] Puenpatom RA, Victor TW. Increased prevalence of metabolic syndrome in individuals with osteoarthritis: an analysis of NHANES III data [J]. *Postgrad Med*, 2009, 121 (6) : 9–20.
- [24] Nieves-Plaza M, Castro-Santana LE, Font YM, et al. Association of hand or knee osteoarthritis with diabetes mellitus in a population of hispanics from Puerto Rico [J]. *JCR: J Clin Rheumatol*, 2013, 19 (1) : 1–6.
- [25] Schett G, Kleyer A, Perricone C, et al. Diabetes is an independent predictor for severe osteoarthritis: results from a longitudinal cohort study [J]. *Diabetes Care*, 2013, 36 (2) : 403–409.
- [26] 王华军, 查振刚. 糖尿病对骨关节炎影响的研究进展 [J]. 暨南大学学报(自然科学与医学版), 2016, 37 (1) : 1–6.
- [27] 吴艳婷, 王颖, 徐谦. 2 型糖尿病合并骨质疏松的研究进展 [J]. *世界最新医学信息文摘*, 2019, 19 (76) : 97–98.

收稿日期:2020-04-16 修回日期:2020-05-12 编辑:王娜娜